MANN-WHITNEY U-TEST

- Non-parametric, no assumptions are made about data fitting a normal distribution
- Used to compare the medians of two sets of data
- It measures the overlap between the two data sets
- You must have between 6 and 20 replicates of data.
- The data sets can have unequal numbers of replicates

Equation:

\[ U_1 = n_1 \times n_2 + \frac{1}{2} n_2 (n_2 + 1) - \sum R_2 \]
\[ U_2 = n_1 \times n_2 + \frac{1}{2} n_1 (n_1 + 1) - \sum R_1 \]

Where:

- **U_1** = Mann Whitney U for data set 1
- **n_1** = Sample size of data set 1
- **\sum R_1** = Sum of the ranks of data set 1
- **U_2** = Mann Whitney U for data set 2
- **n_2** = Sample size of data set 2
- **\sum R_2** = Sum of the ranks of data set 2

You have been provided with data from an investigation into the abundance of Gammarus pulex in pools and riffles of an Exmoor stream. Analyse the data to establish if a.) The pool and riffle areas selected were significantly different from each other and b.) If there is a significant difference of Gammarus pulex between pools and riffles.

Method:

1. Establish the Null Hypothesis \( H_0 \) (this is always the negative form. i.e. there is no significant difference between the sets of data) and the alternative hypothesis \( H_1 \).

   **Velocity \( H_0 \):**
   - There is no significant difference between variable at Site 1 and Site 2
   **Velocity \( H_1 \):**
   - There is a significant difference between variable at Site 1 and Site 2

   **Abundance \( H_0 \):**
   - There is no significant difference between variable at Site 1 and Site 2
   **Abundance \( H_1 \):**
   - There is a significant difference between variable at Site 1 and Site 2

2. Copy your data into the table below as data set 1 and data set 2 and label the data sets.

<table>
<thead>
<tr>
<th>Velocity data</th>
<th>Rank 1 R₁</th>
<th>Rank 2 R₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pools Velocity cm⁻¹</td>
<td>12 12 5 6 14 9 20 8</td>
<td></td>
</tr>
<tr>
<td>Riffles Velocity cm⁻¹</td>
<td>54 55 61 56 31 47 68 54</td>
<td></td>
</tr>
<tr>
<td>Rank 2 R₂</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gammarus pulex abundance</th>
<th>Rank 1 R₁</th>
<th>Rank 2 R₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pools Abundance</td>
<td>27 39 43 2 0 1 3 9</td>
<td></td>
</tr>
<tr>
<td>Riffles Abundance</td>
<td>85 16 80 18 3 5 63 150</td>
<td></td>
</tr>
<tr>
<td>Rank 2 R₂</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

   Helpful hint
   Standard sentences for Mann Whitney \( H_0 \) and \( H_1 \)
   - \( H_0 \): There is no significant difference between variable at Site 1 and Site 2
   - \( H_1 \): There is a significant difference between variable at Site 1 and Site 2
   Use these standard sentences but always state the actual test variable & site names each time
3. Treat both sets of data from the different sites as one data set and rank them all together (i.e. give the smallest value the lowest rank), write the ranks in the shaded boxes.
4. Sum the ranks for each set of data ($\sum R$).
5. Calculate the number of samples in each data set ($n$).
6. Calculate values for $U_1$ and $U_2$.

**Velocity**

$$U_1 = n_1 \times n_2 + 0.5 \times n_2 (n_2 + 1) - \sum R_2$$

$$U_2 = n_1 \times n_2 + 0.5 \times n_1 (n_1 + 1) - \sum R_1$$

**Abundance**

$$U_1 = n_1 \times n_2 + 0.5 \times n_2 (n_2 + 1) - \sum R_2$$

$$U_2 = n_1 \times n_2 + 0.5 \times n_1 (n_1 + 1) - \sum R_1$$

Compare the smallest $U$ value against the table of critical values. If the $U$ value is less than (or equal to) the critical value then there is a significant difference between the data sets and the null hypothesis can be rejected.

7. **Review your result**

**Is there a significant difference?**
- For the velocity in the pools and riffles: Yes / No
- For the Gammarus pulex abundance in the pools & riffles: Yes / No

**Which summary paragraph would you choose out of the following?**

a.) The smallest calculated $U$ value is less than the appropriate critical value ($p = 0.05$), therefore the null hypothesis can be rejected and the alternative hypothesis accepted; there is a significant difference in the velocity of pools and riffles.

b.) Using Mann Whitney U to analyse the data we can be 95% confident that the velocity in the riffles is significantly faster than the pools, however the abundance data for Gammarus pulex showed no significant difference so the null hypothesis must be accepted.

c.) There was no difference in the Gammarus pulex abundance between the pools and riffles.

d.) There was a significant difference in the Gammarus pulex abundance between the pools and riffles.

e.) Using Mann Whitney U to analyse the data we can be 95% confident that the velocity in the riffles is significantly faster than the pools, and that the abundance of Gammarus pulex is significantly higher in the riffles so the null hypothesis can be rejected and the alternative hypothesis accepted.

---

**Critical values for the Mann-Whitney U test** (at the 0.05 or 95% significance level). Critical Values are the point at which the difference between the data sets becomes significant i.e. we are 95% certain that the difference is not due to chance.

<table>
<thead>
<tr>
<th>Value of $n_1$</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>13</td>
<td>15</td>
<td>17</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>7</td>
<td>10</td>
<td>12</td>
<td>15</td>
<td>17</td>
<td>20</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>8</td>
<td>11</td>
<td>14</td>
<td>17</td>
<td>20</td>
<td>23</td>
</tr>
</tbody>
</table>

---

**Which data value/s would you consider to be anomalous and why?**

---

**What type of graph would you use to present this data?**

---